AD-A041 745 PURDUE UNIV LAFAYETTE IND DEPT OF STATISTICS F/G 12/1
SELECTION AND RANKING PROCEDURES: A BRIEF INTRODUCTION.(U)
JUN 77 S S GUPTA N00014-75-C-0455
UNCLASSIFIED MIMEOGRAPH SER-493 NL

1 OF 1
AD
A041745

END

DATE
FILMED
8-77

1
B.S.

See 1473

PURDUE UNIVERSITY

DEPARTMENT OF STATISTICS

DIVISION OF MATHEMATICAL SCIENCES

SELECTION AND RANKING PROCEDURES: A BRIEF INTRODUCTION*

by

Shanti S. Gupta
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #493

June 1977

*This work was supported in part by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University.

SELECTION AND RANKING PROCEDURES: A BRIEF INTRODUCTION*

Shanti S. Gupta

Purdue University

1. ORIGIN OF THE PROBLEM

A common problem faced by an experimenter is one of comparing several categories or populations. These may be, for example, different varieties of a grain, different competing manufacturing processes for an industrial product, or different drugs (treatments) for a specific disease. In other words, we have $k(\geq 2)$ populations and each population is characterized by the value of a parameter of interest θ, which may be, in the example of drugs, an appropriate measure of the effectiveness of a drug. The classical approach to this problem is to test the homogeneity (null) hypothesis H_0: $\theta_1 = \dots = \theta_k$, where $\theta_1,\dots,\theta_k$ are the values of the parameter for these populations. In the case of normal populations with means $\theta_1,\dots,\theta_k$ and a common variance σ^2, the test can be carried out using the F-ratio of the analysis of variance.

The above classical approach is inadequate and unrealistic in the sense that it is not formulated in a way to answer the experimenter's question, namely, how to identify the best category? In fact, the method of least significant differences based on t-tests has been used in the past to detect differences between the average yields of different varieties and thereby choose the 'best' variety. But this method is indirect, less efficient and does not

*This work was supported in part by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University.

easily provide an overall probability of a correct selection. Also the multiple comparison techniques developed largely by Tukey (1949) and Scheffé (1953) arose from the desire to draw inference about the populations when the homogeneity hypothesis is rejected. For details of several multiple comparison techniques, see Miller (1966).

2. SELECTION AND RANKING PROCEDURES

The formulation of a k-sample problem as a multiple decision problem enables the experimenter to anwer his natural questions regarding the best category. Among the early investigators of such procedures are Paulson (1949), Bahadur (1950), Bahadur and Robbins (1950). The formulation of multiple decision procedures in the framework of selection and ranking procedures has been generally accomplished either using the indifference zone approach or the (random sized) subset selection approach. The former approach was introduced by Bechhofer (1954). Substantial contribution to the early and subsequent developments in the subset selection theory has been made by Gupta starting from his work in 1956.

3. DESCRIPTION OF THE TWO APPROACHES

Bechhofer (1954) considered the problem of ranking k normal means. In order to explain the basic formulation, consider the problem of selecting the population with the largest mean from k normal populations with unknown means μ_i, i=1,...,k, and a common known variance σ^2. Let $\bar{x}_i$, i=1,...,k, denote the means of independent samples of size n from these populations. The 'natural' procedure (which can be shown to have optimum properties) will be to select the population that yields the largest $\bar{x}_i$. The experimenter would, of course, need a guarantee that this procedure will pick the population with the largest μ_i with a probability not less than a specified level P*. For the problem to be meaningful P* lies between 1/k and 1. Since we do not know the true configuration of the μ_i, we look for the least favorable configuration (LFC) for which the probability of a correct selection (PCS) will be at least

P*. Since the LFC is given by $\mu_1 = \ldots = \mu_k$, the probability guarantee cannot be met whatever be the sample size n.

A natural modification is to insist on the minimum probability guarantee whenever the best population is sufficiently superior to the next best. In other words, the experimenter specifies a positive constant Δ^* and requires that the PCS is at least P* whenever $\mu_{[k]} - \mu_{[k-1]} \geq \Delta^*$, where $\mu_{[1]} \leq \ldots \leq \mu_{[k]}$ denote the ordered means. Now the minimization of PCS is over the part Ω_{Δ^*} of the parameter space in which $\mu_{[k]} - \mu_{[k-1]} \geq \Delta^*$. The complement of Ω_{Δ^*} is called the indifference zone for the obvious reason. The LFC in Ω_{Δ^*} is given by $\mu_{[1]} = \ldots = \mu_{[k-1]} = \mu_{[k]} - \Delta^*$. The problem is to determine the minimum sample size required in order to have PCS $\geq$ P* for the LFC.

Bechhofer's formulation is more general than what is describ d above. His general ranking problem includes, for example, selection of the t best populations.

In the subset selection approach, the goal is to select a non-empty subset of the populations so as to include the best population. Here the size of the selected subset is random and is determined by the observations themselves. In the case of normal populations with unknown means $\mu_1, \ldots, \mu_k$, and a common variance σ^2, the rule proposed by Gupta (1956) selects the population that yields $\bar{x}_i$ if and only if $\bar{x}_i \geq \max_{i \leq j \leq k} \bar{x}_j - \frac{d\sigma}{\sqrt{n}}$, where $d = d(k,P^*) > 0$ is determined so that the PCS is at least P*. Here a correct selection is selection of any subset that includes the population with the largest μ_i. Thus, the LFC is with regard to the whole parameter space Ω. Under this formulation, for given k and P* we determine d. The rule explicitly involves n. In general, the rule will involve a constant which depends on k, P*, and n. The performance of a subset selection procedure is studied by evaluating the expected subset size and its supremum over the parameter space Ω.

4. TWO EXAMPLES

Example 1. Given five normal populations with unknown means and a common variance 64, it is desired to find which population has the largest mean and to guarantee that the probability of correctly choosing that population is at least 0.90 whenever $\mu_{[5]} - \mu_{[4]} \geq 4$. How many observations must be taken from each population?

We have $\sqrt{n}\,\lambda = 2.5997$ from Table I of Bechhofer (1954) where $\lambda = \Delta^*/\sigma = 0.5$. Thus we take 28 observations from each population.

Example 2. Given the five normal populations as above, it is desired to select a non-empty subset of these populations based on n=8 observations from each population with the guarantee that the population with the largest mean will be included in the selected subset with a probability not less than 0.90. Using Tables in Gupta (1963) (or Gupta, Nagel and Panchapakesan (1973), we find that $d=d(5,.90) = 2.5997$ and $d/\sqrt{n} = .983$. Using a program for generating random normal deviates $N(0 + \alpha\delta, 1)$, $\delta=1$, $\alpha=0,1,2,3,4$, the following sample means $\bar{x}_i$ based on n=7 were observed:

-0.1940 .7987 2.5953 2.8754 4.3841

In this example then, the subset selection rule selects only the observed populations corresponding to observed $\bar{x}_i$ in the interval [3.401, 4.384]. Thus, only the population with the largest $\bar{x}_i$ value is selected in the subset. Note the procedure of Bechhofer will also select the same population with probability of a correct selection equal to .969. The subset selection procedure gives the probability of a correct selection to be between .999 and 1.000 and the associated expected proportions of the number of populations selected in this case lies between .28 and .36.

5. MODIFICATIONS AND GENERALIZATIONS

The above basic formulations have been modified and generalized by several authors. Mahamunulu (1966) has discussed a generalized goal for fixed-size subset selection. His goal is to select a subset of size s from k populations so that the selected subset contains at least c of the t best populations. Of course, the constants c,s,t,k should satisfy some obvious inequalities. The problem is to determine the minimum sample size such that the PCS $\geq$ P* whenever $\mu_{[k-t+1]} - \mu_{[k-t]} \geq d^*$ (in the case of normal means). Desu and Sobel (1968) considered the inverse problem of selecting a subset of the smallest fixed size s given the sample size so that the selected subset will contain the t best of k populations ($t \leq s \leq k$). A formulation for eliminating strictly inferior populations has been used by Desu (1970), and Carroll, Gupta and Huang (1975). Some generalized results in this direction are given by Panchapakesan and Santner (1977).

A restricted subset size formulation has been studied by Santner (1975), and Gupta and Santner (1973). The idea here is to select a subset of random size subject to this size not exceeding a maximum. In the case of k normal populations of unknown means μ_i, $i=1,\ldots,k$, and a common known variance σ^2, let $m(1 \leq m \leq k)$ be the maximum subset size permissible. The goal is to select a subset of size not exceeding m such that the subset contains the population with the largest mean with a probability not less than P* whenever $\mu_{[k]} - \mu_{[k-1]} \geq \delta$. The rule proposed by Gupta and Santner (1973) selects the population corresponding to $\overline{x}_i$ if and only if

$$\overline{x}_i \geq \max\{\overline{x}_{[k-m+1]}, \overline{x}_{[k]} - d\sigma/\sqrt{n}\}$$

where $d > 0$ is a constant to be suitably determined. For given δ, n, and m, the probability guarantee can be met for P* values not exceeding a certain value $P_1 = P_1(k,m,n,\delta)$.

Other generalizations and modifications have been studied by Deverman and Gupta (1969), Sobel (1969), Gupta and Panchapakesan (1972), Alam and Thompson (1973), and Huang and Panchapakesan (1976). There has also been interest in decision-theoretic formulations of subset selection. Specific mention should be made of Studden (1967), Deely and Gupta (1968), Goel and Rubin (1975), Bickel and Yahav (1977), Berger (1977), and Hsu (1977). In the papers by Chernoff and Yahav (1977) and Hsu (1977), Monte Carlo studies have been carried out in the framework of subset selection which show that Bayes procedures can be closely approximated by Gupta type procedures.

6. CONCLUSION

In the last twenty-five years, the research in the area of selection and ranking procedures has progressed steadily and the present count of published papers and technical reports exceeds five hundred. Though these procedures have the potential for application and the use is increasing, it should be admitted that such use is not yet on a large scale. We should perhaps hasten to add that the situation is not unusual considering the fact that it calls for giving up the ingrained habit of testing of hypotheses or tests of significance and ANOVA on the part of applied statisticians. The time is right to remind the theoreticians among ourselves that the gap in the communication with the users is yet to be closed. Some encouraging signs of adopting multiple decision (selection and ranking) theory as realistic alternative to hypotheses testing have again appeared on the horizon. At an international symposium at Purdue in May 1976, there were papers presented on this topic from some competent statisticians, who had earlier not worked on these problems. These include Bickel, Chernoff and Yahav.

ACKNOWLEDGEMENTS

This work was supported in part by the office of Naval Research contract N00014-75-C-0455 at Purdue University. The author wishes to thank Professor S. Panchapakesan and Mr. J. C. Hsu for helpful discussions during the writing of this paper.

BIBLIOGRAPHY

Alam, K. & Thompson, J. R. (1973). A problem of ranking and estimation with Poisson process. Technometrics 15, 801-808.

Bahadur, R. R. (1950). On the problem in the theory of k populations. Ann. Math. Statist. 21, 362-375.

Bahadur, R. R. & Robbins, H. (1950). The problem of the greater mean. Ann. Math. Statist. 21, 469-487. Correction. 22 (1951), 310.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann. Math. Statist. 25, 16-39.

Berger, R. L. (1977). Minimax, admissible, and gamma-minimax multiple decision rules. Mimeo. Ser. No. 489, Dept. of Statistics, Purdue University, West Lafayette, Indiana.

Bickel, P. J. & Yahav, J. A. (1977). On selecting a subset of good populations. Statistical Decision Theory and Related Topics II (Ed. Gupta,S. S. and Moore, D. S.). New York: Academic Press, 37-55.

Carroll, R. J., Gupta, S. S. & Huang, D. Y. (1975). On selection procedures for the t best populations and some related problems. Comm. Statist. 4, 987-1008.

Chernoff, H. & Yahav, J. A. (1977). A subset selection problem employing a new criterion. Statistical Decision Theory and Related Topics II (Ed. Gupta, S. S. and Moore, D. S.). New York: Academic Press, 93-119.

Deely, J. J. & Gupta, S. S. (1968). On the properties of subset selection procedures. Sankhya Ser. A 30, 37-50.

Desu, M. M. (1970). A selection problem. Ann. Math. Statist. 41, 1596-1603.

Desu, M. M. & Sobel, M. (1968). A fixed subset-size approach to a selection problem. Biometrika 55, 401-410. Corrections and amendments. 63 (1976), 685.

Deverman, J. N. & Gupta, S. S. (1969). On a selection procedure concerning the t best populations. Tech. Report. Sandia Laboratories, Livermore, California. Also Abstract Ann. Math. Statist. 40, 1870.

Goel, P. K. & Rubin, H. (1975). On selecting a subset containing the best population - A Bayesian approach. To appear in Ann. Statist.

Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Mimeo. Ser. No. 150. Inst. of Statistics, University of North Carolina, Chapel Hill, North Carolina.

Gupta, S. S. (1963). Probability integrals of the multivariate normal and multivariate t. Ann. Math. Statist. 34, 792-828.

Gupta, S. S. Nagel, K. & Panchapakesan, S. (1973).. On the order statistics from equally correlated normal random variables. Biometrika 60, 403-413.

Gupta, S. S. & Panchapakesan, S. (1972). On a class of subset selection procedures. Ann. Math. Statist. 43, 814-822.

Gupta, S. S. & Santner, T. J. (1973). On selection and ranking procedures - a restricted subset selection rule. Proc. Int. Inst. Statist. 45 (Book 1), 478-486.

Hsu, J. C. (1977). On some decision-theoretic contributions to the problem of subset selection. Mimeo. Ser. No. 491, Dept. of Statistics, Purdue University, West Lafayette, Indiana.

Huang, D. Y. & Panchapakesan, S. (1976). A modified subset selection formulation with special reference to one-way and two-way layout experiments. Comm. Statist. A5, 621-633.

Mahamunulu, D. M. (1967). Some fixed-sample ranking and selection problems. Ann. Math. Statist. 38, 1079-1091.

Miller, R. G., Jr. (1966). Simultaneous Statistical Inference. New York: McGraw Hill Book Co.

Panchapakesan, S. & Santner, T. J. (1977). Subset selection procedures for Δ_p-superior populations. Comm. Statist. This issue.

Paulson, E. (1949). A multiple decision procedure for certain problems in analysis of variance. Ann. Math. Statist. 20, 95-98.

Santner, T. J. (1975). A restricted subset selection approach to ranking and selection problems. Ann. Statist. 3, 334-349.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. Biometrika 40, 87-104.

Sobel, M. (1969). Selecting a subset containing at least one of the t best populations. Multivariate Analysis II (Ed. Krishnaiah, P. R.). New York: Academic Press, 515-540.

Studden, W. J. (1967). On selecting a subset of k populations containing the best. Ann. Math. Statist. 38, 1072-1078.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. Biometrics 5, 99-114.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

14 ~~1. REPORT NUMBER~~
Mimeograph Ser[redacted]-493

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER
9

6 ~~4. TITLE (and Subtitle)~~
Selection and Ranking Procedures: A brief Introduction.

5. TYPE OF REPORT & PERIOD COVERED
Technical rept.,

~~6. PERFORMING ORG. REPORT NUMBER~~
Mimeo . Series #493

10 ~~7. AUTHOR(s)~~
Shanti S. Gupta

8. CONTRACT OR GRANT NUMBER(s)
15 N00014-75-C-0455.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
Purdue University
West Lafayette, IN 47907

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR 042-243

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Washington, DC

11 ~~12. REPORT DATE~~
June 1977

~~13. NUMBER OF PAGES~~
8

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
12 12p.

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Selection and Ranking, Indifference Zone, Subset Selection, Correct Selection, Expected Proportion, Restricted Subset Selection, Modifications and Generalizations.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper presents a brief introduction to selection and ranking methodology. Both indifference zone and subset selection approaches are discussed along with some modifications and generalizations. The paper also mentions two Ph.D. theses of R. L. Berger and Jason Hsu, both students of the author, who have obtained some decision-theoretic results in subset selection.

291 730 4B

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

This article will appear as an introduction to the Special Issue on Selection and Ranking of Communications A-Theory and Methods which should appear later this year and which is being edited by the author.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)